DOCUMENT RESUME

ED 412 253 TM 027 524

AUTHOR Corcoran, Kevin J.; White, Lyle J.; Michels, Jennifer L.; Gilbert, David G.

TITLE Assessing Interrater Reliability of GARF Ratings of Couples' Functioning.

PUB DATE 1997-00-00

NOTE 14p.

PUB TYPE Reports - Research (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Adults; *Diagnostic Tests; Emotional Response; Graduate Students; Higher Education; *Interpersonal Relationship; *Interrater Reliability; Mental Disorders; Problem Solving; Scaling; *Test Construction; Test Reliability

IDENTIFIERS *Diagnostic Statistical Manual of Mental Disorders; Relationship Quality

ABSTRACT

Recently, a great deal of attention has been focused on the development of a system of relational diagnosis to be incorporated into the American Psychiatric Association's diagnostic system, that of the Diagnostic and Statistical Manual (DSM). One of the more intriguing components of this effort is the Global Assessment of Relational Functioning (GARF), which purports to assess the functioning of relational units along dimensions of problem solving, emotional support, and organization. Unfortunately, to date little meaningful reliability data exist concerning this scale. In this study, 64 couples were videotaped while they attempted to resolve their most serious relational problem. Twelve advanced graduate students in training rated the interactions, using the GARF scaling presented in DSM-IV. Results revealed statistically significant, although clinically unsatisfactory, interrater reliability (r=0.43). Results are discussed in terms of the work remaining on the development of the GARF before it is ready for inclusion in a future DSM, and the importance of psychometric properties such as reliability in the development of the GARF and their tests of relational functioning. (Contains 16 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 412 253

Running Head: GARF RELIABILITY

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Kevin Corcoran

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Assessing interrater reliability of GARF ratings of couples' functioning

Kevin J. Corcoran, Ph.D.

Department of Psychology, Southern Illinois University

Lyle J. White, Ph.D.

Department of Educational Psychology and Special Education

Southern Illinois University

Jennifer L. Michels, M.A., and David G. Gilbert, Ph.D.

Department of Psychology

Southern Illinois University

Correspondences concerning this article can be addressed to Kevin J. Corcoran, Department of Psychology, SIUC, Carbondale, IL 62901 (e-mail: corcoran@siu.edu).

TM027524

BEST COPY AVAILABLE

ABSTRACT

Recently, a great deal of attention has been focused on the development of a system of relational diagnosis to be incorporated into the American Psychiatric Association's diagnostic system (DSM). One of the more intriguing components of this effort is the Global Assessment of Relational Functioning (GARF), which purports to assess the functioning of relational units along dimensions of problem solving, emotional support and organization. Unfortunately, to date little meaningful reliability data exist concerning this scale. In the present study, sixty four couples were videotaped while they attempted to resolve their most serious relational problem. Advanced graduate students rated the interactions, using the GARF scaling presented in DSM-IV. Results revealed statistically significant, though clinically unsatisfactory interrater reliability (r = .43). Results are discussed in terms of the work remaining on the development of the GARF before it is ready for inclusion in a future DSM, and the importance of psychometric properties such as reliability in the development the GARF.

Assessing interrater reliability of GARF ratings of couples' functioning

The Diagnostic and Statistical Manual (DSM) system, developed by the American Psychiatric Association, has become the standard diagnostic tool in the field of mental health. This model conceptualizes mental health and personal difficulties as a disorder of an individual. Nevertheless, many therapists recognize that an individual's personal functioning is often influenced by the significant relationships in his or her life. The exclusion of relational functioning from the DSM has historically left marriage and family therapists without the ability to make official diagnoses for disorders evolving from interpersonal relationships, and raising ethical questions for systemic therapists using an individual-based diagnostic system. Recognizing the need to incorporate interpersonal disorders in the DSM, marriage and family therapists formed the Coalition of Family Diagnosis in 1987. The goal of the coalition was to review and stimulate new research in the area of interpersonal disorders, formulate diagnoses for families, and press for the inclusion of such diagnoses in the DSM-IV (Kaslow, 1993). The task of developing a relational functioning scale was difficult given the complexity of identifying and classifying the multitude of variables that influence interpersonal relationships. Compounding this difficulty was the lack of consensus among marriage and family therapists regarding the need for a classification system for interpersonal disorders. Some family therapists, particularly those following a socialist constructivists model, have sighted the fundamental incompatibility between family systems theory and the integration of interpersonal disorders into a manual that has historically focused on the individual (H.A. Goolishian, personal communication, 1989; Strong, 1993). In spite of these difficulties, the Global Assessment of Relational Functioning (GARF) Scale was one of the results of the coalition's efforts. The GARF is designed to allow a clinician to evaluate the degree to

which a family, "meets the affective or instrumental needs of its members" (APA, 1994, p. 758).

The GARF is presented in DSM-IV (APA, 1994) as one of several "axes for further study", but it is analogous to the current Axis V. The multi-axial DSM system reserves Axis V for "judgment of the individual's overall level of functioning" (APA, 1994, p. 30). In the DSM-III-R and DSM-IV, the overall functioning of the individual is assessed with the Global Assessment of Functioning (GAF) Scale. The GAF Scale allows clinicians to rate individuals across three domains, encompassing past and present psychological, social, and occupational functioning. In contrast to DSM categorical assessments, the GAF is based on a 0-100 interval scale, with descriptions of functioning provided at each ten-point interval. Clinicians assign a score between 0 and 100 which best represents an aggregate of an individual's overall functioning in the three domains. The numerical rating that an individual receives is designed to provide information for treatment planning and the prediction of outcome, with lower scores reflecting an increased need for intervention. The GAF is often used in medical and research settings in which there is a need to have a simple diagnostic tool to assess an individual's overall level of functioning prior to and after treatment. Therefore, the GAF has not only served as a treatment indicator but also as a measure of clinical outcome, presumably the GARF is designed to be used in the same way.

The scale is described as easily administered by both experienced and inexperienced staff in facilities such as hospitals, clinics, and public welfare agencies (Kaslow, 1993). Clinicians are permitted to rate a family or relational unit in terms of current and past problem solving, organization, and emotional climate. A statement of functioning pertaining to each of the three areas of focus is provided at each of five 20 point intervals (ranging from "functioning

satisfactorily" to "too dysfunctional to retain continuity" [APA, 1994, p. 758-759]), resulting in a 100 point scale similar to the GAF. Higher scores indicate more optimal functioning.

Because that the GARF is designed for use in multiple settings by staff of varying skill, establishing the reliability of the scale is essential. The limited data available on the reliability of the GAF is of interest due to the fact that the GARF follows the same format as the GAF. Goldman, Skodol, and Lave (1992) summarized the results of studies assessing GAF reliability in an article critiquing Axis V of the DSM-III. Specifically, interclass correlation coefficients of .80 were reported for Axis V ratings of adults using the "joint interview method" (Spitzer & Forman, 1979). Interclass correlations between .57 (Rey, Stewart, Plapp, Bashir, & Richards, 1988) and .61 (Mezzich, Mezzich, & Coffman, 1985) were reported for child and adolescent samples. Inter-rater agreement of 64% between raters of child psychopathology using a 4-point scale version of the Axis V assessment of functioning as a basis for their evaluations (Russell, Cantwell, Mattison, and Will, 1979). Written case summaries have been used as opposed to direct client contact in the studies involving children and adolescents. In addition, test-retest reliability estimates between .49 (Fernando, Mellsop, Nelson, Peace, & Wilson, 1986) and .69 (Spitzer & Forman, 1979) have been reported for clinicians rating adults. Whereas some of these results support adequate reliability for the GAF, the limited variability of clinicians and clients in this relatively small number of studies--especially given that one of the few studies evaluating interrater agreement used a modified form of the GAF (Russell, et. al)--questions the generalizability of these results to other populations.

Although the five 20-point scoring categories that comprise the GARF allow raters to more easily classify relational functioning, a great deal of variability within each 20-point category

is possible. Given the limited reliability information for the GAF and the concern for interrater variability in the GARF, one of the first tasks in the process of validating the utility of the GARF has been to determine interrater reliability. The Group for the Advancement of Psychiatry, Committee on the Family (1996) recently published an article highlighting the development and use of the GARF, information from pilot studies on the instrument, and recommendations for further assessment of the instrument. Of interest to the present investigation are the interrater reliability estimates from three different studies.

In each of three studies, a trainee group, who was instructed on the use of the GARF, rated families presenting with concerns ranging from marital discord and child misbehavior to problems with sexual and physical abuse. A trainee group in Rochester, New York and second group in Commerce, Texas observed and rated sets of families through one-way mirrors. Details on the training of the raters was omitted. Twenty families were assessed by 52 raters to gain 106 GARF ratings in Rochester. Twenty-eight families were assessed by 9 raters to gain 80 GARF ratings in Commerce. The reported results described interrater reliability as significant at $p = .02$ for both the Rochester, New York and Commerce samples. A third trainee sample in Montreal was comprised of 30 mental health professionals who watched and rated two video tapes of family interviews. Comparisons among raters in this study found that interrater reliability was significant at $p = .0000$ [sic]. In all three studies, however, no information about the degree and directionality of the relationship among the raters' scores was provided. Some measure of size of the effect observed, as recommended by the American Psychological Association (1994), would be more appropriate. For example, with a sample of 30, a correlation coefficient of roughly .45 would result in a significance level of $p = .01$. Statistically significant, but hardly an acceptable

interrater reliability--minimum interrater reliability is generally agreed upon as .80 or 80% (Kazdin, 1982). In sum, the manner of presenting results relevant to interrater reliability chosen by Committee on the Family (1996) is of limited use in assessing the utility of the GARF.

Dausch, Milowitz, and Richards (1996) were more specific in their evaluation of the interrater reliability of the GARF. Raters were trained to rate family interactions with the use a manualized form of the GARF. A criterion rater provided training in family systems theory and the use of the GARF to three bachelors-level psychology students who had little prior experience with families. Following the training, each rater independently rated a videotaped interactions subset of the sample of 73 (41 women and 32 men) bipolar patients and their families during a problem solving discussion. Intraclass correlation coefficients between the criterion rater and each of the three trained raters were: $r = .81$ for rater 1 (46 families), $r = .94$ for rater 2 (72 families), and $r = .81$ for rater three (51 families). Each of the three correlations was significant at $p < .001$. Dyadic pairings of the three B.A.-level raters averaged a level of agreement of .72 ($p < .001$).

Although the previous intraclass correlations involing extensively trained individuals fall within an acceptable range, the generalizability of these ratings to other groups that have not had systematic training is questionable. This is especially relevant given Kaslow's (1993) assertion that the GARF is easily administered by both experienced and inexperienced staff. Thus, the interrater reliability of the instrument when used by those without extensive formal training or those facing real world time constraints remains in doubt. The present study investigated the interrater reliability of the GARF with raters who have not experienced systematic training with this instrument. The intent of this study is to provide information relevant to the utility of the scale

in settings that it is likely to be used.

METHOD

Participants

The participants in this study were 32 couples from research conducted at the University of Florida and 32 couples from research conducted at Southern Illinois University. Mean age of participants was 30.96 years (SD=11.27), and they were married for an average of 6.78 years. Couples were paid for their participation.

Procedure

Participant procedure.

As part of a larger study, couples completed personality and relationship questionnaires. Included in this packet was the Couple's Problem Inventory (Gottman, 1979) in which participants rated the severity of ten major marital problems (from "no problem" to "highly severe problem"). A researcher attempted to determine which problem was rated as most troublesome. If both partners rated the same problem area most severe, the couple was asked to discuss that problem. If there was disagreement, the researcher read the top three rated problems of each spouse aloud and the couple agreed on a problem both felt to be distressing. Spouses were asked to independently write about the problem in 100-130 words, writing as though they were directly communicating with their spouse. One of the spouses was randomly assigned to read his/her problem description to the other spouse who was instructed to remain silent during the reading. Following the problem presentation, the couple was given ten minutes to discuss and attempt to resolve the conflict. They were instructed to come to a real solution that satisfied both their needs. The problem solving discussion was videotaped by a camera mounted on a tripod across

the room from the couple, recording a full body view of both spouses.

Rater procedure

Raters were twelve graduate student clinicians in Clinical Psychology and Counselor Education in at least their second year of graduate training. All participants had experience with individual counseling/psychotherapy. None had any formal practicum training in family therapy. Each rater was given a reproduced copy of the full DSM-IV description of the rating scale to be used for the GARF (pages 758 and 759 of DSM-IV). They were asked to study the coding system, watch each video (and replay sections if necessary), provide an overall rating of relational functioning on a scale of 0 to 100--as the GARF is designed to be used. Because more specific ratings generally result in stronger reliabilities, the raters were also asked to provide ratings for each of the three area scored by the GARF: Problem solving; Organization; and Emotional Climate.

RESULTS

Overall, the average GARF score for the 64 couples was 66.88 (SD=16.71). Reliability on the GARF ratings was computed using a Pearson Product moment correlation, yielding an r of .43 ($p < .001$). Interrater reliabilities were also computed for the three subareas of the GARF, resulting in correlation coefficients of .43, .41, and .37 for "emotional climate", "organization", and "problem solving", respectively (all $p < .005$).

DISCUSSION

The present results paint a disappointing picture of the reliability of the GARF as a measure of relational functioning in a brief dyadic interaction. Although the interrater reliability was statistically significant, it falls far short of practical significance--suggesting less than 20%

shared variance between raters viewing the same tape.

At the outset, it is important to keep in mind that these ratings are based on brief interactions outside of a therapeutic context, and should be seen as preliminary. Nevertheless, the raters had the opportunity to review the tapes several times before making their ratings, and they were not unexperienced in assessing clients. At a minimum, these data suggest that more work is needed before the GARF is moved from the appendix to the main portion of a future edition of the DSM. This should not be surprising given the concerns about the reliability of the GAF which served as the basis for the GARF.

The use of an assessment of relational functioning along a continuum from positive to problematic does not seem as controversial as a typological diagnostic system for relational problems, especially given the inclusion of "competent and healthy functioning" (Group for the Advancement of Psychiatry, Committee on the Family, 1996, p. 156). It is impressive that this group attempted to take into account the concerns of all potential users of the GARF. Nevertheless, it is disappointing to see the ease with which the GAP Committee on the Family so easily dismisses the importance of psychometric factors, "[the] GARF is not, and perhaps never will be, a psychometrically refined instrument for finely honed research use" (p. 159). Presumably, this orientation influenced the decision to report interrater reliabilities in a less meaningful way (i.e., reporting only significance levels rather than percentages of agreement, correlation coefficient, or other measures of effect size). Acceptable interrater reliability seems to be a minimal psychometric requirement for a rating scale. Furthermore, interrater reliability with highly trained raters does not approximate real world usage of DSM-IV. What does a GARF rating mean if it is not reliable among practitioners likely to use it? Clinicians ought to be as

concerned about the ethics of using an unreliable scale as they are about using terminology with which they disagree.

At this point, what directions should researchers and clinicians consider? One possibility is to sever the link with the GAF and move to a 5 point scale. Our raters found the 100 point scale confusing; several suggested five quintiles (which already have descriptions in DSM-IV). If family clinicians are willing to accept the premise that family systems theory is compatible with DSM-based diagnoses--historically focused on individual psychopathology--then researchers ought to continue to strive to satisfy the need for diverse health care providers to communicate using a common language, while not overlooking the requirements of good scale construction. As the Committee on the Family of the GAP notes, the GARF has been proposed, "not as a final and definitive answer to the complex issues in the assessment of relational systems, but perhaps a step forward" (1996, p. 169). These results suggest that the GARF has begun to frame important questions for family clinicians. It is the centrality of the questions which makes the importance of fundamental psychometric issues such as interrater reliability more not less essential. As a result, more work lies ahead before researchers and practitioners approach the answers to those questions.

REFERENCES

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (3rd ed.)*. Washington DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders (3rd ed., revised)*. Washington DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington DC: Author.

American Psychological Association (1994). *Publication manual of the American Psychological Association (4th edition)*. Washington, D.C.: Author.

Dausch, B. M., Miklowitz, D. J., Richards, J. A. (1996). Global assessment of relational functioning scale (GARF): II. Reliability and validity in a sample of families of bipolar patients. *Family Process 35*: 175-189.

Fernando, T., Mellsop, G., Nelson, K., Peace, K., & Wilson, J. (1986). The reliability of axis V of the DSM-III. *American Journal of Psychiatry 143*: 752-755.

Goldman, H. H., Skodol, A. E., Lave, T. R. (1992). Revising axis V for DSM-IV: A review of measures of social functioning. *American Journal of Psychiatry 149*: 1148-1156.

Gottman, J. M. (1979). *Marital interaction: Experimental investigations*. New York: Academic Press.

Group for the Advancement of Psychiatry, Committee on the Family. (1996). Global assessment of relational functioning scale (GARF): I. Background and rationale. *Family Process 3:*, 155-172.

Kaslow, F. (1993). Relational diagnosis: Past, present, and future. *The American Journal of Family Therapy 21*: 195-204.

Kazdin, A.E. (1982). *Single case research designs: Methods for clinical and applied settings.* New York: Oxford.

Mezzich, A. C. Mezzich, J. E., & Coffman, G. A. (1985). Reliability of the DSM-III vs. DSM-II in child psychopathology. *Journal of the American Academy of Child Psychiatry 24:* 273-280.

Roy, J. M., Stewart, G. W., Plapp, J. M., Bashir, M. R., Richards, I. N. (1988). Validity of axis V of DSM-III and other measures of adaptive functioning. *Acta Psychiatry Scandinavia 77*: 534-542.

Russel, A. T., Cantwell, D. P., Mattison, R., and Will, L. (1979). A comparison of DSM-II and DSM-III in the diagnosis of childhood psychiatric disorders: multi axial feature. *Archives of General Psychiatry 36*: 1223-1226.

Spitzer, R. L., & Forman, J. B. W. (1979). DSM-III filed trials: Initial experience with the multi axial system. *American Journal of Psychiatry 136*: 818-820.

Strong, T. (1993). DSM-IV and describing problems in family therapy. *Family Process 32*, 249-253.